View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Hertfordshire Research Archive

Unsupervised Learning-based Anomalous Arabic Text Detection

Nasser Abouzakhar, Ben Allison, Louise Guthrie
NLP Research Group
Dept of Computer Science
The University of Sheffield
E-mail: N.Abouzakhar@dcs.shef.ac.uk, B.Allison@dcs.shef.ac.uk, L.Guthrie@dcs.shef.ac.uk

Abstract

The growing dependence of modern society on the Web as a vital source of information and communication has become inevitable. However, the Web has become an ideal channel for various terrorist organisations to publish their misleading information and send unintelligible messages to communicate with their clients as well. The increase in the number of published anomalous misleading information on the Web has led to an increase in security threats. The existing Web security mechanisms and protocols are not appropriately designed to deal with such recently developed problems. Developing technology to detect anomalous textual information has become one of the major challenges within the NLP community. This paper introduces the problem of anomalous text detection by automatically extracting linguistic features from documents and evaluating those features for patterns of suspicious and/or inconsistent information in Arabic documents. In order to achieve that, we defined specific linguistic features that characterise various Arabic writing styles. Also, the paper introduces the main challenges in Arabic processing and describes the proposed unsupervised learning model for detecting anomalous Arabic textual information.

1. Introduction

Consider the following problem: given a collection of documents, how does one determine whether any text(s) within the collection are out of context with the collection as a whole? This paper presents work which addresses this problem automatically for the case where the documents are written in Arabic. Some previous work (Guthrie et al., 2007) has approached this problem where the documents are written in English; however, the characterisation of English text is something well-studied and comparatively well-understood. Arabic, on the other hand, has received little attention within the NLP community (and until recently, almost none) and provides a much more challenging application for an already ambitious task. However, we believe that this paper shows that in certain cases, anomalous documents can be detected with high accuracy, and the framework and methods we propose can easily be extended to improve performance still further.

We call a text which is out-of-context with the rest of a collection an *anomaly*. We appreciate that this subjective notion of anomaly embraces texts that are out-of-context because their topic is different, or they have different authors, different writing styles and genres, and so on, or indeed any combination of the above.

As a language, Arabic has attracted much attention in recent years for various socio-political reasons and due to links between certain groups in the Middle East such as Alqaeda and terrorism. However, much technology is extremely naïve, performing simple keyword-spotting and leaving more complicated analysis to humans. However, with the wealth of intelligence being gathered exceeds the human capacity to analyse, and the gap between what can be analysed and what is being collected is undoubtedly increasing. Thus the need for more sophisticated automatic techniques to pre-process text is great; this paper proposes a framework which at least partially fulfills that goal. Figure 1 shows a high level model and indicates the task of detecting anomalous text. The task model is scanning through data collection to identify any inconsistency.

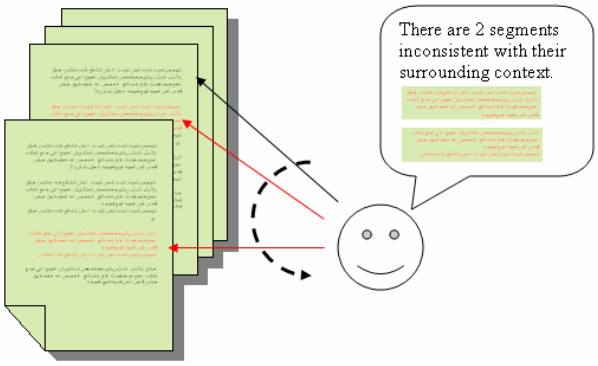

Figure 1: The task model

As a highly inflected language Arabic poses several challenges in terms of language processing in general and anomalous text detection in particular. To explore those challenges, this research proposes a frame work of unsupervised learning solution that is able to locating unusual segments of text or simply collection of documents that are used abnormally. Our work centers around collecting various content and performing the tasks for developing detection models that can aid NLP community in better understanding Arabic language processing and analysis.

2. Related Work

In general, the problem of detecting anomalous Arabic text could be related to other similar problems such as authorship analysis, automatic plagiarism detection or

detecting breaches of intellectual copyright (Clough et al., 2002). However, in the artificial intelligence community, similar problems to ours have been approached differently such as detecting anomalous communication traffic between various computer networks (Abouzakhar and Manson, 2002). The major difference between those various applications is the type of knowledge that is extracted from the relevant environment. This knowledge is evaluated in order to select and derive specific features which are used to differentiate what could be abnormal object/event from its normal boundaries.

The problem of detecting anomalous Arabic text has not been previously explored by the NLP community. However, similar little work on this problem has been performed by various research centres. Abbasi and Chen (2005) adopted machine learning classifiers for authorship analysis by extracting linguistic features from online Arabic messages. These features are evaluated for patterns of terrorist communications. They developed a multilingual model by incorporating a message extraction component tailored toward online messages. Our work investigates three unsupervised learning approaches which assume no prior knowledge about what could be normal language. These techniques have already been applied in a previous successful work (Guthrie et al., 2007) for detecting anomalous English textual information. However, English stylistic features are different to Arabic but similar unsupervised learning models could be applied by gearing these models toward the Arabic unique characteristics.

3. Major Challenges in Arabic

Arabic poses various specific challenges in terms of the language writing stylistic properties and rules. However, the linguistic complexities of Arabic language create serious issues for the anomaly detection. Also, Arabic (De Roeck and Al-Fares, 2000; Rozovskaya et al., 2006) has morphological characteristics that pose several critical problems to text analysis techniques. These specific Arabic language characteristics impact the anomaly detection parameters such as the language lexical and stylistic features and part of speech features … etc.

For example, the absence of the diacritics which are optional in Arabic could lead to an ambiguous case. It is impossible to distinguish between the words حُب /hubb/love and حَبَ/habb/seed without using the relevant diacritics. The lack of diacritics in most of the modern standard Arabic (MSA) is considered as a major challenge to many of Arabic NLP tasks. Recent studies (Maamouri et al., forthcoming) have indicated the impact of diacritisation in text-based NLP research in general and syntactic analysis and parser development in particular. As a highly inflected language, Arabic constructs its vocabulary through a complicated derivational process using root words. These sorts (Habash, 2006) of linguistic characteristics pose challenges in terms of features extraction, morphological and text analysis.

3.1 Segmentation Ambiguity

One of the features that we extract from our Arabic corpus is the percentage of conjunctions in each paragraph. The conjunction "و" "wa – and" has got certain properties that limit the process of its extraction. One property of this conjunction is that is should not be joined to the following word. Also, this conjunction could be used as a normal letter like any other letter that can be used with any possible word. However, if it is part of a word then we would not expect to have a space between "و" as a letter and the following letter of the same word. However, people tend to ignore including a space between "و" as a conjunction and the following word. Figure 2 shows an example of a segmentation ambiguity during the extraction of the conjunction "و" "wa – and".

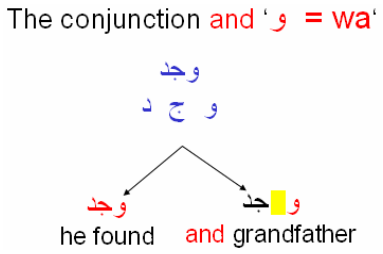

Figure 2: Segmentation Ambiguity

Figure 2 indicates how "و" as a letter is used as part of a word that means "he found" where there is no a space between "و" and the following letters. Also, it shows how "و" is used as a conjunction to the following word which gives the meaning of "and grandfather". Notice that we have to have a space in the latter case. So, in terms of this feature extraction we base our decision on the following rule: if there is no space we can not decide whether this conjunction is present or not therefore, we decide not to count as a conjunction if ambiguous.

3.2 Abridged Words Ambiguity

This kind of ambiguity often leads to inflectional ambiguity as well. Figure 3 shows an example of abridged word ambiguity. The example shows an Arabic sentence "وسنقولها" "and we will say it" that is written using a single Arabic word. This sentence contains the abridged form of the pronoun "نحن" "We". Another feature that we extract is the percentage of pronouns in each paragraph. However, the abridged form of the pronoun "نحن" is represented by one letter "ن" "N". Of course such a letter could be found in any possible Arabic word.

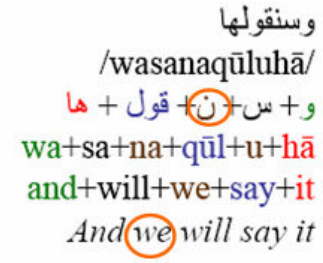

Figure 3: Abridged Word Ambiguity

Therefore, if the pronoun "نحن" is abridged we can not decide whether it is present in such a sentence. So, we the decision is not to count a pronoun if ambiguous. Also, notice how the root word say changed its form when it is used as part of the whole sentence.

4. Method

For each document in a collection, we characterise the document as a vector of features whose elements are defined. Then, for each vector, we calculate its dissimilarity from each other vector in the collection. Thus for each vector we produce a set of scores (as many as there are documents in the collection, minus one), and the anomaly score for that segment is the sum of all these scores.

We employ three different measures of (dis)similarity, namely the cosine similarity widely used for IR and city block distance and chebychev distance measure.

Let the vectors *a* and *b* represent the two vectors in question. Then the cosine of the angle between the two vectors is:

$$s\left(\vec{a},\vec{b}\right)=\frac{\sum_{i=1}^{n}a_i b_i}{\sqrt{\sum_{i=1}^{n}a_i^2}\sqrt{\sum_{i=1}^{n}b_i^2}}$$

And the dissimilarity between *a* and *b* is 1 - *s(a,b)*. Similarly, the *city block distance* between the two vectors is simply:

$$d\left(\vec{a},\vec{b}\right)=\sum_{i=1}^{n}\left|a_i-b_i\right|$$

Where in each case a_i is the i^{th} element of the vector *a*.

The Chebychev (dissimilarity) distance between the two vectors *a* and *b* is the maximum distance between both vectors. The distance between ***a***=(***a1, a2***, etc.) and ***b***=(***b1, b2***, etc.) vectors is computed using the formula:

$$Max_i=\left|a_i-b_i\right|$$

where $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are the values of the *i*th element at vectors ***a*** and ***b***, respectively.

In some senses, the city-block distance between the two vectors is the more simple measure, as it is simply the sum of the absolute values of the differences. Furthermore, the cosine similarity measure is widely accepted in the field of IR; however, we believe that the city block distance provides a useful basis for comparison given that the application is quite difference from information retrieval.

In all cases, the vectors *a* and *b* are normalised so that each feature is on the same scale. For this work, we pick a simple normalisation to z-scores. For the i^{th} element of a vector, let s_i represent some segment *s*'s score for feature *i* and μ_i represent the mean of the i^{th} feature across all documents in the collection. Similarly, let σ_i equal the standard deviation for the i^{th} feature. Then the normalised z-score for s_i, which we have thus far called a_i, is:

$$a_i=\frac{s_i-\mu_i}{\sigma_i}$$

5. Experiments & Evaluation

Our assumption is that the language of any out-of-context text would have a minority occurrence with respect to the whole normal document and hence suggests that it is anomalous. In terms of detecting this anomalous text, the whole document could be represented as a one space with four possible measures of text anomalousness, as shown in figure 4. The 4 possible measures for identifying any particular text as anomalous are as follows:

- Detected & Anomalous (DA): correctly identified anomalous text and is called True Positive (TP)
- Detected & Not anomalous (DN): incorrectly identified text as anomalous whereas in fact is normal and is called False Positive (FP)
- Not detected & Anomalous (NA): incorrectly identified text as normal whereas in fact is anomalous and is called False Negative (FN)
- Not detected & Not anomalous (NN): correctly identified normal text and is called True Negative (TN)

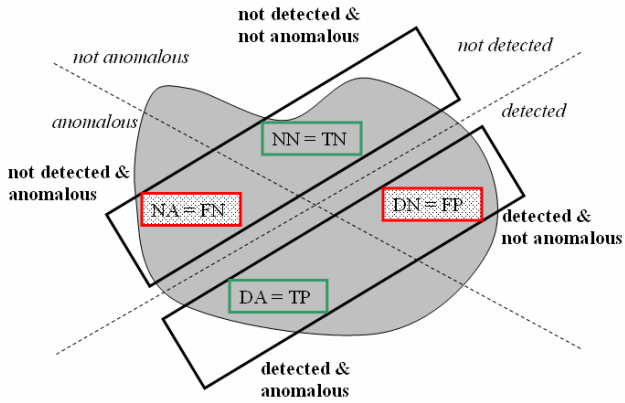

Figure 4: Measures based on Text Anomalousness

We incorporated a feature extraction component to allow the use of a more specific language features tailored specifically toward Arabic. Various experiments for evaluating the developed models indicated a high level of success in terms of detecting anomalous Arabic textual information. Here we are primarily concerned with applying statistical analysis solutions to identifying anomalous Arabic text. This process begins by

determining and extracting the specific linguistic features of the language, however, this task (Smrz, 2005) is complicated by the requirements of Arabic language processing and lack of effective and applicable resources such as morphological analysers. Similarly, the linguistic complexities of Arabic inflectional morphology create issues for the developing syntactic features and semantic annotation.

Various Arabic Web sites were used for text collections including Aljazeera.net as the main sources for normal text. Three main types of text were used to represent anomalous segments, religious description text, social text and novels, as shown in Table 1. Aljazeera.net is a well established news sources in the Middle East that reflect the news from its TV broadcast channels. The collected news articles including those inserted out-of-context information would then be segmented and statistical features were computed for each segment.

We defined Arabic specific linguistic features that characterise various writing style, however, well-established general stylistic features have been applied as well. For example, general stylistic features include:

- average word length
- average sentence length
- average number of question sentences
- frequent words
- vocabulary richness
- short sentences
- long sentences
- punctuation marks
- prepositions
- conjunctions
- personal pronouns
- demonstrative pronouns
- relative pronouns
- positive words
- negative words
- …

It is interesting to make use of Arabic specific features as they can effectively differentiate various writing styles. For example, religious description text tends to refer to Quran in which diacritics are often used. However, the "general" features such as average size of sentences in a segment and so on are also useful. For example, Arabic specific features include:

- diacritics
- diphthongs
- nunation
- elongation
- numbers (Arabic and/or Indian)
- calendar (lunar and/or crescent (qamari or Hijri))
- …

The underlying assumption is that if two segments have dissimilar writing style based on selected features, both should have a high dissimilar score i.e. high distance measure. Each segment has been converted into a vector of the features and in a form of numerical values. Three unsupervised learning approaches were used, the cosine (dis)similarity (distance) measure, city block distance and chebychev distance have been used to measure the distance of each segment (paragraph) from every other segment in the same document.

Type of text articles	Number of articles	Number of words (each article)
Aljazeera.net news	50	200 - 500
Alarabiya.net news	50	200 - 500
Religious	30	200 - 500
Social	30	200 - 500
Novels	30	200 - 500

Table 1: The Corpus data

In order to appreciate the differences amongst all features in terms of their effects on the distance measures we normalize our vectors by calculating the Z-score of each feature individually before applying the distance measures. All three measures represent three solution models that used the numerical vectors of each segment in the document. This research work is operating within the context of two domains the first drawn from the machine learning community and the other from the natural language processing community. It investigates the deployment of unsupervised learning as a statistical approach to identify anomalous Arabic text. This is to detect as well as predict future textual anomalies, and hence minimise their negative effects. Consequently, it may provide innovative solutions that can be implemented in a cost-effective manner.

6. Results

In these experiments various types of text were used to represent normal and anomalous segments. In each experiment one segment of those various anomalous sources (religious, social and novels) was randomly selected and inserted into 50 of Aljazeera.net news segments. These news segments were used as a normal text. In each experiment one anomalous article is inserted into 50 of the normal news articles and a stylistic vector that represents all features is derived for each segments. The cosine (dis)similarity measure, city block distance and chebycheve distance measures are calculated and recorded separately. This process is repeated 30 times for each type of anomalous documents. The outcome in both cases is a ranked list of scores for each segment according to how anomalous it is with respect to all other segments in the same document. Those segments that are not likely to belong to the whole collection would have high scores and are considered as anomalous text.

The evaluation process that is used to assess the performance of the developed models has been concentrated on conducting various experiments for detecting and predicting anomalous text. The Top-n method is used to measure the performance of the developed models in terms of their detection rate. This method indicates how many times each anomalous segment appeared in the top n of the ranked list. The generated lists of ranked scores are used to evaluate the performance of each model. The results in figures 5, 6 and 7 show how often the various anomalous segments are ranked in the Top 1, Top 3 and Top 10. The results indicate the number of truly detected anomalous segments (DAs or TPs) which represent the detection rates. All figures compare both solution models of the cosine (dis) similarity, city block and chebychev distance measures with random selection/detection by chance. Figure 5 shows the results of inserting 30 religious text articles one by one into 50 of Aljazeera.net news articles. Due to the likely heavy use of diacritics in many of the religious text, all three detector models are able to separate most of religious segments (~90%) in the top 10. All results indicate that the city block is quite similar to chebychev, however both distance measures outperform cosine similarity measure.

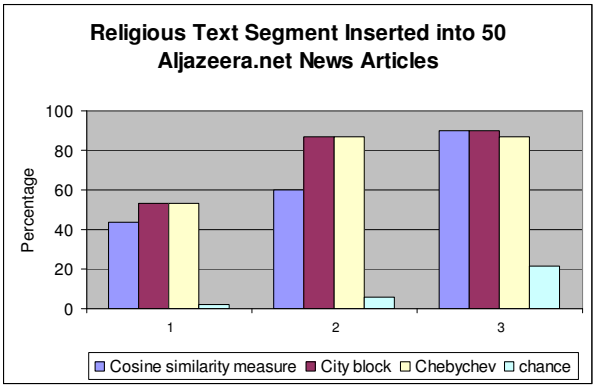

Figure 5: Religious text is used as anomalous

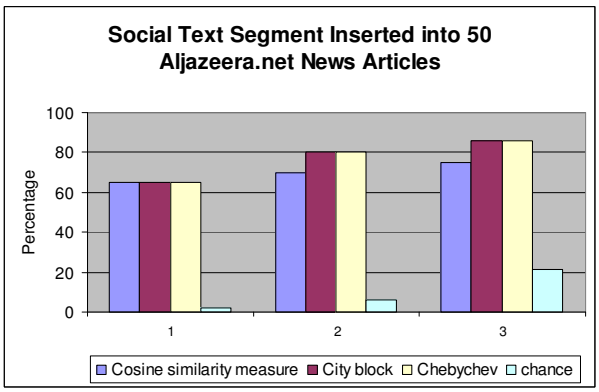

Figure 6: Social text is used as anomalous

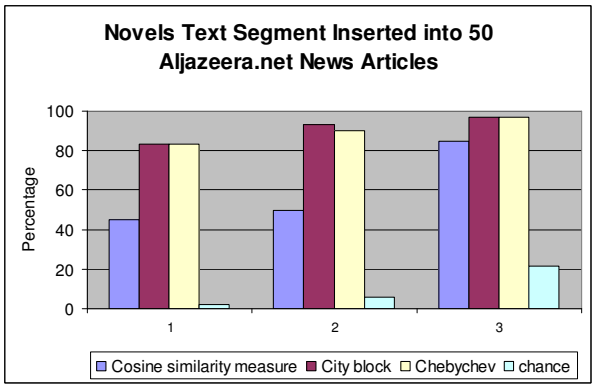

Figure 7: Novels text is used as anomalous

Previous work (Guthrie et al., 2007) in English textual documents have applied the same unsupervised learning approaches and had reached similar results in terms of having city block distance outperforms cosine similarity measure.

7. Conclusions

This paper introduced the problem of detecting anomalous Arabic text by automatically extracting linguistic features from various documents and evaluating those features for patterns of inconsistent information. In order to achieve that, we defined specific linguistic features that characterise various Arabic writing styles. Three unsupervised learning approaches were used to develop anomalous detection models, the cosine (dis)similarity (distance) measure, city block distance and chebychev distance measures have been used to measure the distance of each segment (paragraph) from every other segment in the same document. Our experiments showed encouraging results and our research indicated a significant unsupervised learning power in the application of anomalous Arabic text detection.

By means of evaluation, as well as empirical evidence, we are able to determine the effectiveness of the developed models and assumptions. The performance of the three developed detection models has been evaluated and the results indicate that the city block is quite similar to chebychev, however both distance measures outperform cosine similarity measure. However, all models have achieved a significant increase in the detection rates (> 75%) in terms of the detected anomalous segments in the Top 10.

8. References

Abouzakhar, N. and Manson, G. (2002). An Intelligent Approach to Prevent Distributed Systems Attacks, The Journal of Information Management and Computer Security, 10 (5): 203 - 209.

Al-Sughaiyer, I. and Al-Kharashi, A. (2004). Arabic Morphological Analysis Techniques: A Comprehensive Survey, Journal of the American Society for Information Science and Technology, 55 (3): 189-213.

Abbasi, A. and Chen, H. (2005). Applying Authorship

Analysis to Extremist-Group Web Forum Messages, Intelligent Systems, IEEE Computer Society.

Abbasi, A. and Chen, H. (2005). Applying Authorship Analysis to Arabic Web Content, Springer-Verlag, Berlin Heidelberg.

Al-Shartuni, R. (2006), Elementary Arabic Morphology 1 (Mabadi Al-Arabiyyah). Translated by: Hamid Hussein Waqar.

Ben Amara, N. E. and Bouslama, F. (2003). Classification of Arabic Script Multiple Sources of Information: State of the Art and Perspectives, International Journal on Document Analysis and Recognition (IJDAR), Springer-Verlag.

Clough, P. Gaizauskas, R. Piao, S. and Wilks, Y. (2002). METER: MEasuring TExt Reuse. In proceedings of the 40th Anniversary Meeting for the Association for Computational Linguistics (ACL-02), pp.152-159, 7-12 July, University of Pennsylvania, Philadelphia, USA.

De Roeck, A. N. and Al-Fares, W. (2000). A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots, Proceedings of Association for Computational Linguistics (ACL oo).

Guthrie, D. Guthrie, L. Allison, B. and Wilks, Y. (2007). Unsupervised Anomaly Detection, International Joint Conferences on Artificial Intelligence, India.

Habash, N. Arabic Tutorial, (2006). The fifth international conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy.

Maamouri, M. Kulick, S. Bies, A. (2006). Diacritisation: A Challenge to Arabic Treebank Annotation and Parsing. The Challenge of Arabic for NLP/MT, International Conference at the British Computer Society (BCS), London.

Maamouri, M. Bies, A. and Kulick, S. (2006). Diacritisation in Arabic Treeback Annontation and Parsing, University of Pennsylvania

Messoudi, A. Lamel, L. and Gauvain, J. (2004). The LIMSI RT-04 BN Arabic System, Proceedings of the EARS RT-04 Workshop.

Rozovskaya, A. Sproat, R. and Benmamoun, E. (2006). Challenges in processing colloquial Arabic, The Challenge of Arabic for NLP/MT, International Conference at the British Computer Society (BCS), London.

Smrz, O. (2005). Introduction to Arabic Natural Language Processing, Lecture notes, Dept. of Middle Eastern Studies, The University of West Bohemia, Pilsen, Czech Republic.